Journal of Biomedical Informatics **34,** 85–98 (2001)
doi:10.1006/jbin.2001.1012, available online at http://www.idealibrary.com on IDEAL®

Eneida A. Mendonça,[1] James J. Cimino, Stephen B. Johnson, and Yoon-Ho Seol

Department of Medical Informatics, Columbia University, New York, New York 10032

Received December 7, 2000; published online June 12, 2001

The large and rapidly growing number of information sources relevant to health care, and the increasing amounts of new evidence produced by researchers, are improving the access of professionals and students to valuable information. However, seeking and filtering useful, valid information can be still very difficult. An online information system that conducts searches based on individual patient data can have a beneficial influence on the particular patient's outcome and educate the healthcare worker. In this paper, we describe the underlying model for a system that aims to facilitate the search for evidence based on clinicians' needs. This paper reviews studies of information needs of clinicians, describes principles of information retrieval, and examines the role that standardized terminologies can play in the integration between a clinical system and literature resources, as well as in the information retrieval process. The paper also describes a model for a digital library system that supports the integration of clinical systems with online information sources, making use of information available in the electronic medical record to enhance searches and information retrieval. The model builds on several different, previously developed techniques to identify information themes that are relevant to specific clinical data. Using a framework of evidence-based practice, the system generates well-structured questions with the intent of enhancing information retrieval. We believe that by helping clinicians to pose well-structured clinical queries and including in them relevant information from individual patients' medical records, we can enhance information retrieval and thus can improve patient-care.

[1]To whom correspondence should be addressed at Department of Medical Informatics, Columbia University, 622 West 168th Street, Vanderbilt Clinic, 5th Floor, New York, NY 10032. Fax: (212) 305-3302. E-mail: mendonca@dmi.columbia.edu.

I. INTRODUCTION

Several studies have assessed the needs of clinicians for access to information pertinent to clinical practice [1–3]. With the large and rapidly growing number of information sources relevant to health care, an increasing number of professionals and students are gaining free access to an expanding volume of information that was previously inaccessible (e.g., full-text journals, patient education materials, computer-assisted instruction). Seeking and filtering useful and valid information can be difficult because of the speed with which the information is accumulating and the increasing number of biomedical information sources [4]. Keeping up to date with the advances in medical science and incorporating evidence to make safe and accurate diagnostic, therapeutic, and management decisions is a difficult task [5]. The development of decision support tools designed to provide relevant and current evidence to clinicians is a possible solution to this problem [6]. Such tools include those that facilitate access to, extraction of, and summarization of evidence.

In this paper, we describe a model that builds on the several different, previously developed techniques to identify information themes that are relevant to specific clinical data. We add a framework of evidence-based practice to these methods, and build a system that generates well-structured questions with the intent of enhancing information

1532-0464/01 $35.00

retrieval (IR). Additional algorithms and methods have been developed to identify data in the electronic medical records (EMR) that pertain to individual patient care. We begin this paper by reviewing studies of information needs of clinicians and describing principles of information retrieval. We also examine the role that standardized terminologies can play in the integration between a clinical system and literature resources, as well as in the IR process. Finally, we describe the new model and the implementation of a system based on this model.

II. BACKGROUND

Although many workers in health-care settings (e.g., physicians, nurses, dieticians, pharmacists) have information needs, the majority of studies have examined information needs and IR related to physicians. Information retrieval systems are considered to be valuable tools for practicing physicians [7]; however, study results show that physicians still have difficulty using such resources [8]. During clinical practice, physicians and other health-care workers see patients whose problems raise questions that clinicians can answer by doing literature retrieval [9, 10]. An online information system that would conduct searches based on individual patient data could both have a beneficial influence on the particular patient's outcomes and educate the health-care worker [10]. Researchers have thus explored ways to integrate information resources into clinical systems. They have tried to improve bibliographic search results, as well as to use clinical data to trigger retrievals, by studying the process of human thinking and specific information needs.

By IR, we mean the process of retrieving documents from computer-based information resources. Based on Salton's work [11], Hersh and colleagues [12] divide the problem of IR into four processes: (1) indexing, (2) query formulation, (3) retrieval, and (4) evaluation and refinement. Traditional health-related IR systems represent and store their set of documents in a unique database, and use particular expressions to state information needs in terms of queries (Fig. 1A). The rise of Web technologies and the need to make clinical evidence easily available to clinicians contributed to the development of digital libraries, where materials converted to digital format are published, and different digital collections are linked in a way transparent to the end user. This expansion (Fig. 1B), however, raises questions related to technical (e.g., equipment, logon procedures, access), conceptual (e.g., information needs, knowledge representation, terminologies), and organizational connections (e.g., copyright, agreement between institutions) [13]. Sections II.A, II.B, and II.C present a more detailed discussion of these questions, while Section II.D describes work in evidenced-based decision making that can be drawn on in a possible solution to these questions.

II.A. Information Needs and Information Retrieval

Knowledge of the nature and scope of the information needs experienced by clinicians is essential for the design and implementation of automated methods capable of obtaining and managing clinically relevant information [14]. Lancaster and Warner [15] describe three basic types of information needs: (1) the need to solve a certain problem or make a decision; (2) the need for background information; and (3) the need to keep up with the latest information for a given subject. The interaction of users with an information system varies based on these different needs [16].

Researchers have used a variety of techniques to understand what types of questions physicians generate, and what resources physicians use to answer these questions. Some studies assumed that needs exist, and identified and measured those needs (questions physicians ask, urgency, etc.) [1, 2, 5, 17, 18]. Other studies examine what information physicians used to answer their questions. Some of the latter studies also looked at computer-based sources [1, 5].

Results of these studies showed that physician information needs are highly specific to patients' problems, and many studies categorized needs into classes such as *diagnosis* or *treatment*. Primary care has been the area that most of these studies examined. In general, questions on treatment are the most common, followed by questions on diagnosis and etiology. Ely and collaborators [18] found that the most frequent questions asked by family physicians were of the form of "What is the cause of symptom X?," "What is the dose of drug Y?," and "How should I manage disease or finding Z?" Covell [1] found that the questions were not generalized, but rather were practice related. Thus, a possible question would be "What is the dose of digoxin for a patient with heart failure and associated renal impairment?" rather than "What is the dose of digoxin?"

Only a few studies have examined nurses' information needs [19, 20]. Corcoran-Perry and Graves [19] found that most of the information sought by cardiovascular nurses was related to patient-specific data (general nursing care, medication administration, laboratory reports, etc.), followed by institution-specific data and domain knowledge (nursing knowledge and knowledge from related disciplines). In the Corcoran-Perry and Graves study, patient care

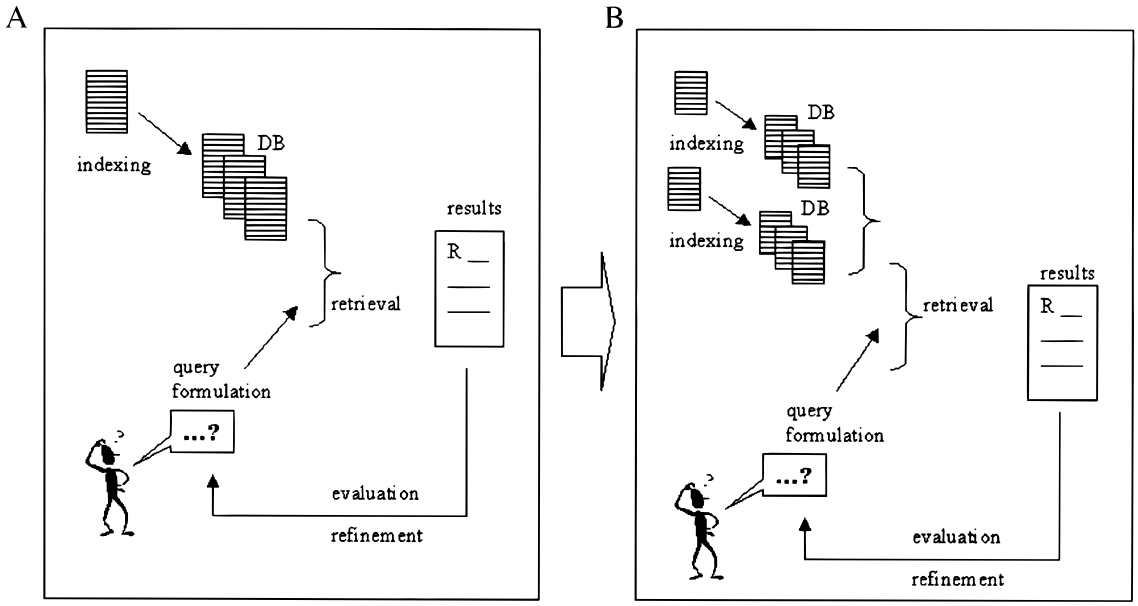

FIG. 1. (A) The traditional approach of searching a single information resource. (B) The new environment encountered in a "digital library."

(assessing, planning, giving, and managing direct patient care) accounted for 76% of the reasons for seeking information.

Several studies also addressed the use of online information sources [1, 5, 21–26]. According to these studies, the regular use of computers for literature searching is minimal. Studying the online access of physicians and medical residents from hospital units (academic and community settings), ambulatory clinics, office-base practices, and a clinical pharmacy group, the researchers found that online resource usage was only a few times a month. A systematic review by Hersh and Hickam [27] concluded that IR systems have had a modest but important influence in health care, but that there are many unanswered questions about how well they are used. For example: why does the overall use of IR systems occur just 0.3 to 9 times per physician per month, if physicians have 2 unanswered questions for every 3 patients? Would improvements in technology increase the usage of IR systems? Do the IR systems contain the appropriate information? Are the searchers retrieving all of the potential relevant material from a given topic or question?

Addressing the issue of information quality, Gorman [28] looked at the quality of information (whether the quality of information was sufficient for application in clinical practice), and found that only one-third of retrieved papers contained "high-quality" evidence. Sackett and colleagues [29] studied the feasibility of finding and applying evidence during clinical rounds by using an "evidence cart" that contains multiple sources of information. Of all searches, 81% sought evidence that could affect diagnostic or treatment decisions. From the successful searches, 25% led to a new diagnostic skill, an additional test, or a new management decision, and 23% corrected a previous clinical skill, diagnostic test, or treatment. Smith [30] analyzed 13 studies of physicians' information needs, including several described here, and concluded that the physician's information tool of the future might be a combined electronic patient record and Internet resource.

II.B. Terminology Issues

In general, health literature uses two types of index terms: controlled terminology (or thesaurus) and raw words (usually using each word in the document, or part of the document, as an indexing term). Indexing can be done manually (human indexing) or automatically. Controlled terminologies contain preferred terms, synonyms, and relations between terms (hierarchical, or other relationships). Examples of terminologies used in health-care literature are the Medical Subject Headings (MeSH) [31] by the National Library of Medicine, and the Cumulative Index to Nursing and Allied Health Literature—CINAHL—Subject Headings. Retrieval

is achieved through the formulation of a search based on a question posed by a user. The information need is translated (by the user) into a question (query), which contains the indexing terms, Boolean operators (AND, OR, NOT), and other advanced searching operators (e.g., subheadings, explosion operation, and proximity operators). Two aspects related to terminology should be considered: (a) the translation of clinical terms to a target form (recognized by the desired information resource), and (b) the type of query posed to the information source (e.g., therapy query).

Due to the heterogeneous sources of information available (e.g., textbooks and bibliographic databases), the use of different terminology is a significant challenge to searching for the best answer. Much work has been done on ways to translate clinical terms to a controlled terminology. Manual translation between terminologies is very time-consuming, requiring automated tools to support the process. Although some work on translation was done for other purposes (e.g., expert systems, clinical information systems), it has potential applicability in IR systems. Some studies focus on the translation of clinical terms in programs that assist in medical diagnosis, using synonyms [32, 33], word stem algorithms [32, 34], and spelling checkers [33]. Other studies focus on the translation of free text into MeSH terms, using frequency of words in the title [22], word stemming, spelling checking, and key word synonyms.[35] A few studies applied several of these techniques [36–38]. Others applied techniques to translate free text into SNOMED terms [39]. All methods mentioned above were interactive, helping the user to select the terms. Automated translation using lexical [40] and morphosemantic [41] algorithms has also been studied, as well as the use of a semantic network [42], and a frame-based system for mapping among terminologies [43]. Finally, the Unified Medical Language System (UMLS) [44] has been used by many researchers as a means of mapping between existing vocabularies.

There are, however, limitations to these approaches. Maintenance is problematic, especially when dynamic terminologies (such as MeSH) are involved [42]. Automated methods can potentially facilitate the translation and maintenance processes, especially if formal definitions are created. More recent studies have made clear the need for concept-oriented terminologies to describe clinical encounters, support data reuse, and data comparison across different representations [45–48]. Description logic-based languages, such as KRSS [49] and GRAIL [50], have been developed for the management and processing of concept-oriented terminologies. To be able to support comparisons of data represented using different terminologies, these languages must support unambiguous concept representation [48].

IR can be a complex strategy that includes, in addition to search terms, information such as which fields should be searched, what the clinical task is, and Boolean, truncation, or proximity operators. Understanding the major clinical research types (e.g., therapy, diagnosis, etiology and causation, natural history and prognosis, and economics) and the basic methodologies associated to each research type can help in effectively retrieve material of value when making clinical decisions [51]. For example, clinical trials are frequently used to evaluate therapeutic or prevention interventions, while cohort studies are often associated with natural history and prognosis studies. In searching for articles on therapy, for example, a user may include search terms such as “clinical trial” or “blinded.” It is important to understand how the different resources index articles of each type of research, and how this indexing can be used in retrieving information. Haynes and colleagues [52] suggested that search terms (e.g., “randomized controlled trial,” “cohort studies,” and “specificity”) that select studies that are “at the most advanced stages of testing for clinical application” can be a potential method for improving the detection of studies of high-quality data.

II.C. Linking Clinical Information to Online Resources

The idea of bringing together clinical data, medical knowledge, and expert systems to assist patient care is not a recent conception. The promotion of Integrated Advanced Information Management Systems (IAIMS) was an important move toward the development of such integration. Also significant were digital library projects originated by with the National Foundation of Science in 1994 [13]. Researchers have suggested that bibliographic information should be integrated with clinical applications, especially electronic medical record systems, to facilitate clinicians’ access to scientific evidence, clinical practice reports and guidelines, as well as to other decision-support tools. In this way, information retrieved is personalized based on the context of patient individual characteristics [10, 53].

Applications have varied from a simple integration between clinical and bibliographic systems, allowing the user to access the retrieval system and select the desired information to be retrieved from the clinical system (e.g., Medical Desktop [54], Meta-1 Front End [55]), to more complex systems, which use patient record or clinical reports to anticipate the user’s needs (e.g., Hepatopix [56], Psychtopix [57], Chartline [58], IQW[59], the Medline Button [60], and Infobuttons.[61]).

II.D. Principles of Evidence-Based Practice

Evidence-based medicine (EBM) has been defined as the "conscientious, explicit, and judicious use of the current best evidence in making decisions about the care of individual patients" [62, 63]. Evidence-based medicine focuses on questions related to the central tasks of clinical work: diagnosis, etiology, prognosis, therapy, prevention, and other clinical and health care issues. EBM requires the ability to access, summarize, and apply information from the literature to day-to-day clinical problems. This requires an understanding of the structure of medical literature and the use of clinical filters in searching medical databases. The process involves four basic steps: (1) converting information needs into focused questions, (2) efficiently tracking down the best evidence with which to answer the question, (3) critically appraising the evidence for validity and clinical usefulness, and (4) evaluating performance of the evidence in clinical application.

The first step for any search is to define the question that needs to be answered. It involves identifying a question that is important to the patient's well-being, is interesting to the physician or health care provider, and that he/she is likely to encounter on a regular basis in his/her practice [64]. This requires the understanding of the information needs and a careful definition of the question, and it is often more difficult than first anticipated [51]. Sackett and colleagues describe this step as the formulation of a "well-built clinical question" [65]. In practice, a well-built question is usually of two types: (1) general knowledge or "background" questions, and (2) patient-specific knowledge questions on diagnosis, treatment, prognosis, etc. (referred to as "foreground questions"). Background questions have two essential components: a question root (why, who, what, where, how, when) followed by a verb, and a disorder or aspect of a disorder. A foreground question has at least three of four elements: (a) a patient or problem being addressed, (b) an intervention, (c) a comparison interventions (optional), and (d) an outcome of interest. A fifth element, the type of clinical work (or where clinical questions arise from) is also important in the process of IR.

III. THE MODEL

The foundation of the model we describe is the integration of a query-enhancement module (which uses evidence-based principles) with an electronic medical record within an interface to a digital library (see Fig. 2). The model (shown in Fig. 3) has three major components: *display*, *data*/*knowledge*, and *processing*. The model supports systems that facilitate search, presentation, and summarization of online medical literature.

Display. The display component is essential for the interaction between the user and processes that run in the system. The display supports three different interfaces: (1) a clinical interface (clinical information system), (2) a query interface (where search strategies are displayed and data entry is allowed), and (3) a summarization interface (where data retrieved from the bibliographic collections are displayed and summarized and data entry is permitted to refine strategies).

Data/*knowledge.* Specific databases and KBs are needed to support the processes. The extraction process is supported by a medical knowledge base. This KB contains the clinical terminology and knowledge needed to support the use of patient data for the various processes, including the extraction of information relevant to specific patient problems, the discovery of new knowledge from the patient's data, and the application of this knowledge in the process of selecting appropriate questions to ask. The KB encompasses concepts derived from clinical settings, such as those used in laboratory, pharmacy, clinical descriptions (findings and

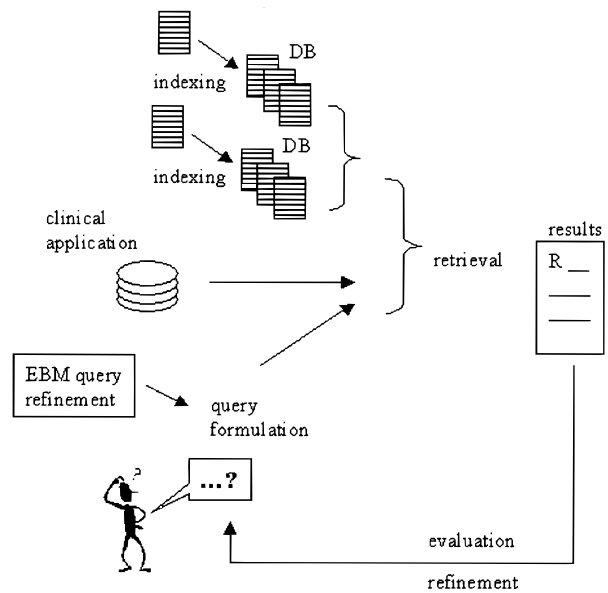

FIG. 2. The influence of our proposed model on the digital library environment.

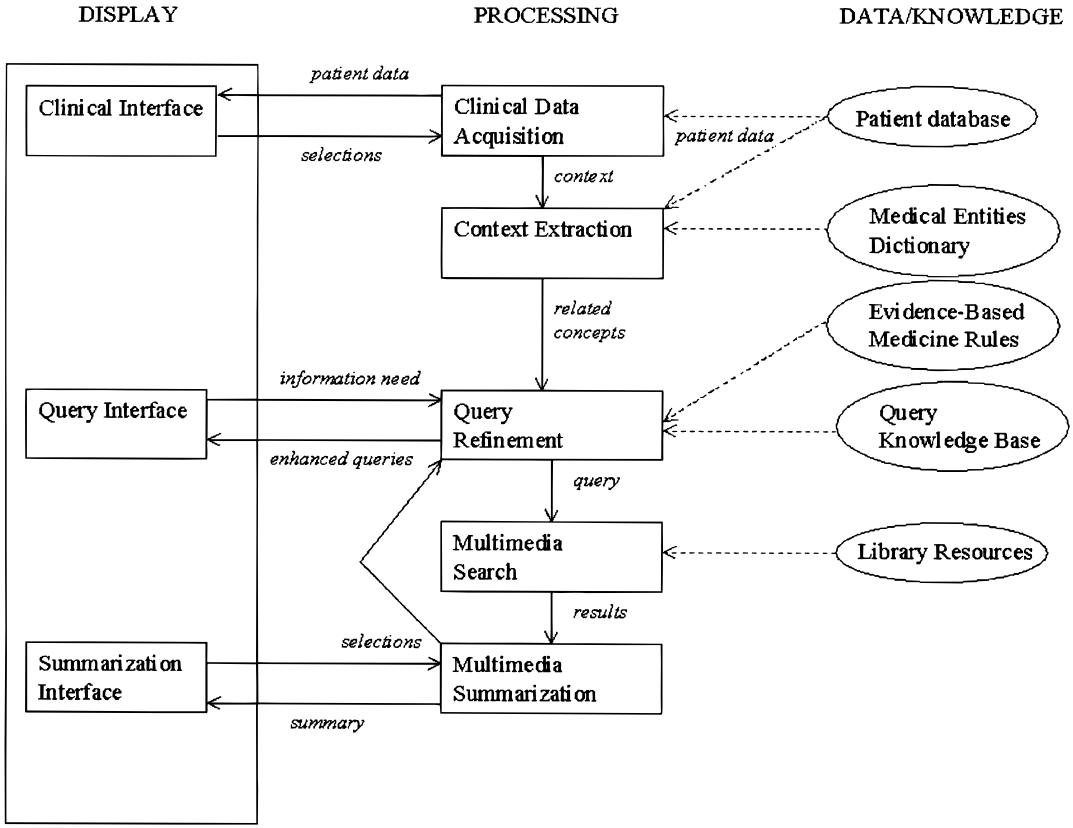

FIG. 3. Integrating the clinical system with information sources.

symptoms), and diseases, among others. It also contains information on the relationships among these concepts. These relationships may be of different types such as "is–a," "is–part-of," and "is–caused-by." For instance, the KB may contain information about the circulatory system, defining heart diseases as diseases of the circulatory system, and defining cardiac enzymes as a laboratory test related to heart diseases.

A model of information needs, based on clinical generic queries and evidence-based medicine principles, is used by the query refinement process to generate queries that are related to an individual patient. This model contains generic queries based on questions that are frequently asked by clinicians in daily practice.

The library data encompass medical knowledge resources available in electronic form, such as medical textbooks, clinical journals, health literature databases, online information, and evidence-based medicine resources.

Processing. The processing component comprises five subprocesses: clinical data acquisition, context extraction, query refinement, multimedia search, and multimedia summarization. These subprocesses are responsible for the integration of views, data, and KBs. Clinical data acquisition is a process that retrieves patient data from a clinical database or electronic medical record. Context extraction is responsible for the collection of relevant information from the same database. By relevant we mean information that is related to the clinical situation of the patient.

This process assesses the information retrieved by the data acquisition process for display to user and, using information from a KB, retrieves related data from the same clinical database. The result is a set of clinical concepts that are directly related to a particular patient. This set is then used to guide the IR.

The query refinement process is responsible for query matching and refinement. This process uses all information acquired by the two previously described processes and a KB of generic queries to find the queries that are more

appropriate for that particular set of patient data. The matching process is based on the clinical data extracted, particularly patient data extracted from the screen the user is examining. As the query is defined, the multimedia search process selects the appropriate library resources. The multimedia summarization process is responsible for the presentation of the information retrieved from the literature resources [91, 92].

III.A. Implementation Methods

To illustrate the model, we describe the methods we have developed and are implementing in our institution. The description focuses on the clinical application, the extraction of relevant clinical data from the medical record, the query construction, and the KBs.

Previous techniques. Researchers in our institution previously developed techniques to identify information that is relevant to specific clinical data, establishing direct links between data the user is examining in an electronic medical record and potential questions they might have. The first technique involved the development of a set of *generic queries* by analyzing questions posed by clinical users to establish common semantic and syntactic patterns [66]. For example, "Is [laboratory test] indicated in [disease or syndrome]?," "What is the drug of choice for [disease or syndrome]?," and "Is [pharmaceutical component] indicated in [disease or syndrome]?" The Medline Button [60] was the first implementation of this approach. This application used the UMLS Metathesaurus to translate patient discharge diagnosis and procedures codes from ICD9-CM [67] to MeSH term, assembling these terms into search statements, and passing them on to a MEDLINE search engine.

The second technique developed was the "infobutton" [61]. Infobuttons use generic queries that exploit the hierarchical and semantic links in the Medical Entities Dictionary (MED) [68], a knowledge-based terminology, to identify additional terms that can be used in the generic questions. The MED organizes concepts into a semantic network of frame-based term descriptions. The relationships in the network provide definitional knowledge about the individual terms. The infobutton uses information about the identified concepts to select the relevant generic queries. Furthermore, it traverses the links in the MED to find terms that are appropriate for searching. For example, instead of searching for "penicillin-sensitivity test" (which may be not productive in certain bibliographic databases), it searches for "penicillin." Examples of generic queries triggered by this process are "What are the indications for [antibiotic]?," "What is the toxicity of [antibiotic]?," and "What is the mechanism of action of [antibiotic]?" The semantic and hierarchical links in the MED contain the information necessary to support this process. Infobuttons are currently implemented in two major applications at New York Presbyterian Hospital: PatCIS (a Patient Clinical Information System that allows patients to review data from their own medical records) [69], and WebCIS (Web-based Clinical Information System for clinicians) [70].

The infobutton applications identify relevant terms to build patient-oriented queries consisting of the relationships between concepts in the MED and the logic of making valid connections. For instance, to connect myocardial infarction with CK-MB (laboratory test) requires knowledge of the relationships between myocardial infarction (MI) and heart diseases, and of relationships among intravascular CK test, creatine kinase, cardiac enzymes, and heart disease. Figure 4 shows how a concept is linked to related clinical data [90].

The system now being developed makes significant enhancements in several areas: extracting what is known about the individual patient whose record the clinician is reviewing, matching to an improved model of user information needs, and enabling the user to refine the query.

Content extraction. In the current implementation, the display component allows users to interact with the system using three different interfaces: clinical, query, and summarization. The initial interaction is done in the clinical interface, e.g., when a physician is examining a patient's medical record. Clinical data are either text data (e.g., reports, discharge summaries) or coded (e.g., laboratory test results such as glucose). The user can access the library environment from any part of the clinical information system. At this point, a series of processes take place. The context extraction process uses a KB, the MED, to guide the extraction of relevant concepts from the patient's medical record. Relevant concepts are data that may be related to the clinical situation of the patient, the patient demographics, or the content of the screen displayed. For example, the laboratory result screen may display a blood glucose level of 54. The KB knows that blood glucose level of 54 is low and extracts hypoglycemia as a relevant concept. A general natural language text processor, MedLEE [71], helps the system to identify clinical concepts in the text reports of the medical record (e.g., discharge summaries). Figure 5 shows examples of data extracted from the medical record.

We used two methods for extracting content from the medical record:

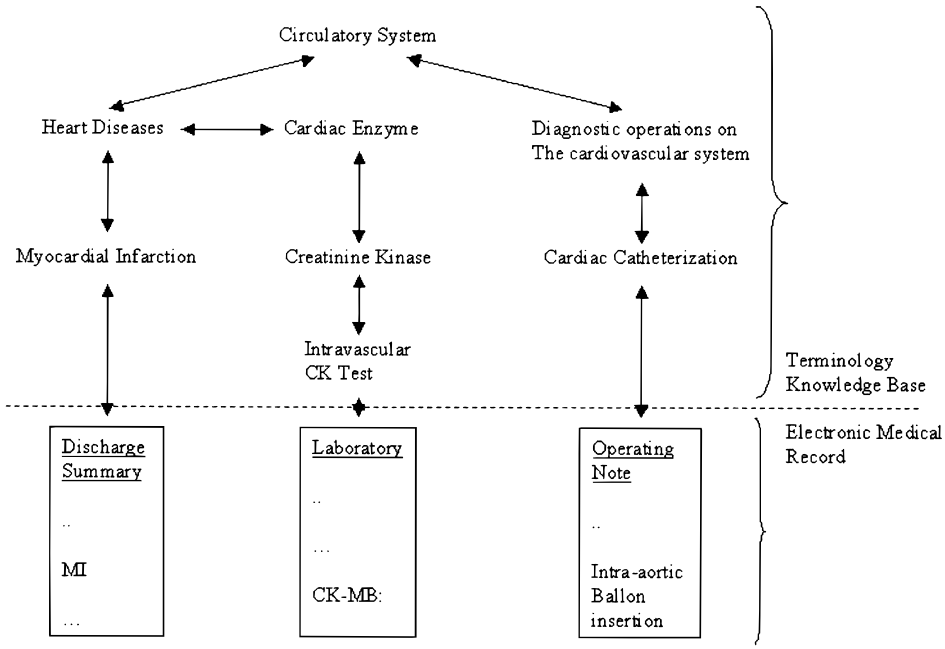

FIG. 4. Linking medical concepts to patient data. (Adapted from Zeng [90].)

—Using co-occurrence of MeSH terms in MEDLINE citations in association with the search strategies optimal for evidence-based medicine to automate construction of a KB [72]

—Using the information stored in the MED: literal attributes, hierarchical links, and semantic links.

The first method adds to the MED information extracted from medical literature. Data stored in the MED allow us to link concepts to retrieve other possible relevant concepts. In an additional example, the system captures from the latest discharge summary the information that this patient has congestive heart failure; it then uses the semantic links in the MED to suggest the treatments of interest. Congestive heart failure, for example, is linked in the MED to diuretics and cardiac drugs through the semantic link "has-related-pharmaceutical-chemical." Using the hierarchical links, the system finds two cardiac drugs (enalapril preparations and captopril preparations—ACE inhibitors). In the MED, the "main-MeSH" attribute gives us the synonym term in MeSH, Taking the "enalapril preparations" as the example, the attribute value returns "enalapril." All information acquired by the content extraction process is represented as conceptual graphs [73] and stored in two forms: application context (information captured from the screen the user is examining) and clinical context (other information retrieved from the medical record). Figure 6 shows how the data extracted are represented as conceptual graphs.

Query definition and refinement. This information is then passed on to the query refinement process. This process uses a new model of information needs and performs two major tasks: query matching and query refinement [74]. The model is based on the taxonomy of generic clinical questions developed by Ely and colleagues [75]. Questions that do not have all components of a "well-built" clinical question defined by Sacket and others [76] were slightly modified to make them well-built. Examples of questions in this representation are shown in Fig. 7.

Questions are represented internally as conceptual graph and are assigned a score according to the frequency with which they were observed in Ely's study and the degree to which they match the information extracted from the medical record. The first matching process uses the application context information and finds the closest generic queries available, such as "What causes ⟨disease/condition⟩?," "What is the best treatment for ⟨disease/condition⟩? ⟨drug therapy/procedure 1⟩ ⟨drug therapy/procedure 2⟩ . . . ⟨drug therapy/procedure *N*⟩," and "What is the efficacy of ⟨drug therapy/procedure⟩ for ⟨disease/condition⟩?" Semantic types and relations are based on the semantic net of the UMLS.

All possible matching questions are displayed to the user

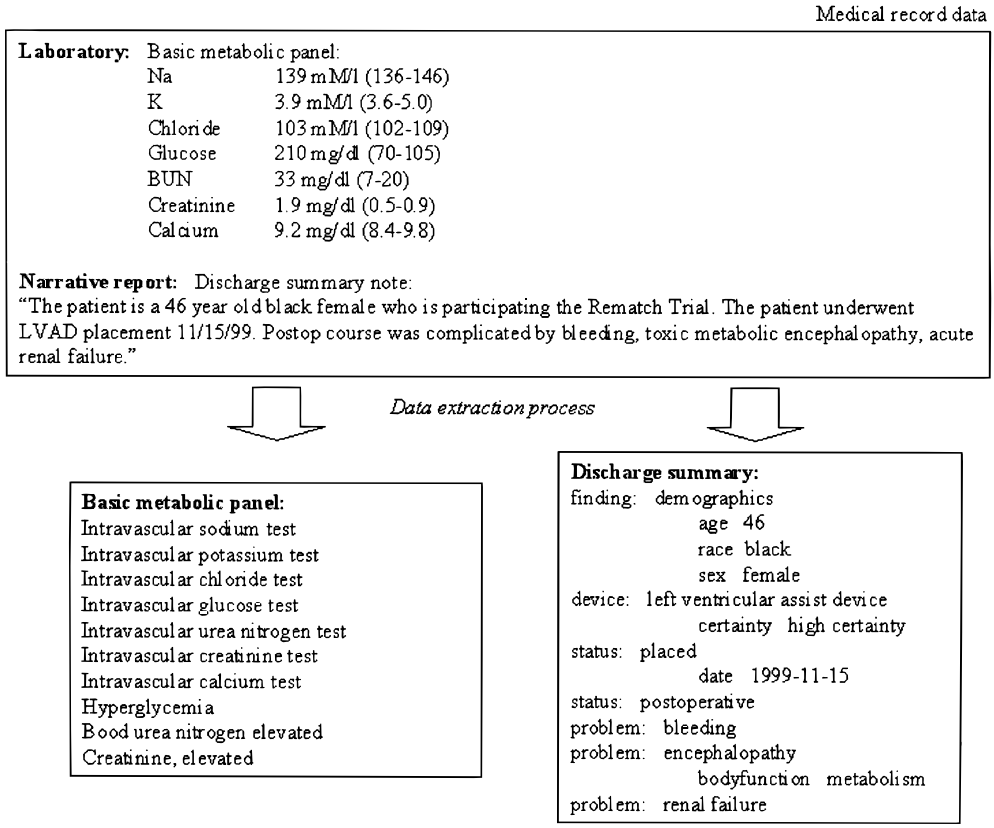

FIG. 5. Partial view of data from a medical record, and the information collected after the data extraction step.

by the query interface. If necessary, refinement can be done by using additional information from the clinical context. A new set of questions is then presented. The query chosen by the user is passed to the multimedia search process. For example, if the initial questions presented by the system are related to the treatment of heart failure in general, but the user is not satisfied, the refinement process adds clinical context information (e.g., patient is 60 years old) to the query. A bibliographic search on medical databases, which are more likely to have the answer for a specific therapy question, such as EBM Reviews—Best Evidence, retrieves five articles, four of them on the use of enalapril with or without diuretics, and one on the use of captopril and diuretics. The information is presented to the user who decides to refine a bit more using more information from the medical record: patient age group. The results come down to one article on the "short-term survival with for myocardial infarction."

The current system is able to process demographic, laboratory, microbiology, and narrative reports data presented in the medical records. The MED contains some 67,000 concepts, including about 206,000 synonyms, 100,000 hierarchic relationships, 167,000 other relationships, and 138,000 mappings to other terminologies, such as the UMLS. The generic queries' database contains about 180 generic questions divided in 50 groups.

IV. DISCUSSION

Studies have suggested that the use of computer-based IR systems by clinicians enhances the quality of patient care [7] by encouraging better use of evidence in the development of care plans [65], and by helping clinicians to keep up with the health literature [77]. Making evidence easily available at the point of care increases the extent to which clinicians seek evidence and incorporate it in their practice [29].

The practice of evidence-based medicine requires retrieval of relevant information or evidence to support clinical decisions. The information retrieved must be appropriate to the

Generic semantic representation of a microbiology laboratory culture and sensitivity test:

[LPRO]->(AE)->[ANTB]->(DS)->[PFUN]->(PO)->[ORGM]<-(PP)<-[OATT]

Interpretation: A procedure assesses the effect of an antibiotic which disrupts a physiologic function which is a process of an organism which has an attribute (sensitive/resistant).

Example: Culture & Smear Site
Specimen description: catheterized urine.
Culture: >100K col/ml Morganella Morganii
Organism: Morganella Morganii
Method: Microscan MIC.
Sensitivity test: Ampicillin 2S, Sulfamethoxazole R, Cephalexin 2S (partial result)

Conceptual graph representation of the test:

[LPRO: Culture & Smear Site]->(AE)->[ANTB: ampicillin]->(DS)->[PFUN: undefined]->(PO)->
[ORGM: Morganella Morganii]<-(PP)<-[OATT: sensitive]
[LPRO: Culture & Smear Site]->(AE)->[ANTB: sulphamethoxazole]->(DS)->[PFUN: undefined]->(PO)->
[ORGM: Morganella Morganii]<-(PP)<-[OATT: resistant]
[LPRO: Culture & Smear Site]->(AE)->[ANTB: cephalexin]->(DS)->[PFUN: undefined]->(PO)->
[ORGM: Morganella Morganii]<-(PP)<-[OATT: sensitive]

FIG. 6. Conceptual representation of a microbiology laboratory culture and sensitivity test.

care of a specific patient. The optimization of IR strategies for clinical relevance and the integration of infrastructure building blocks to support the context-specific retrieval and application of evidence in practice are major challenges in the development of informatics infrastructure for evidence-based practice [78]. The model we describe in this paper supports the integration of clinical systems with online information sources, making use of information available in the EMR to enhance searches and IR. Our objective is to facilitate retrieval by automatically generating queries that account for specific characteristics of individual patients.

1. What causes congestive heart failure?
Question root + verb: what + cause
Condition: congestive heart failure
Clinical task (category): diagnosis

2. What is the efficacy of enalapril for congestive heart failure?
Patient/condition: congestive heart failure
Intervention: enalapril
Comparison: none
Clinical outcomes: morbidity, mortality
Clinical task (category): therapy

3. What is the best treatment for heart failure? ACE inhibitors or diuretics?
Patient/condition: congestive heart failure
Intervention: ACE inhibitors
Comparison: diuretics
Clinical outcome: morbidity, mortality
Clinical task (category): therapy

FIG. 7. Representation of well-built questions.

We use patient information to guide clinicians in finding evidence to use in their patient's care. For example, a nurse examines a patient record and observes that the patient is not compliant with his medication. A possible generated question is "Which is more effective in improving medication compliance in patients with ⟨disease/syndrome⟩: ⟨procedure⟩ or ⟨procedure⟩?" This structured question can be filled in using the information extracted from either the medical record or the KB. A final question is modified to "In an elderly patient, which is more effective in improving medication compliance in patients with congestive heart failure: educational interventions or reminder devices?" This enables us to perform queries that address specific information needs such as the questions observed by Covell and colleagues (What is the dose of digoxin for a patient with heart failure and associated renal impairment?).

One challenge we face is to decide which information we should extract from the medical record to enhance the queries for IR. We built a KB that, in conjunction with the semantic information in the MED, guides the extraction process. We are using the information in MEDLINE citations and the

UMLS in this KB. The extraction method and preliminary evaluation data are described in detail elsewhere [72]. A pilot study demonstrated that the extraction method was suitable for literature retrieval, especially for data related to the clinical task "therapy."

In such construction, terminology plays an important role, especially in the process of integrating the information resources with the clinical systems. Terminology issues arise in the mapping of terms from the UMLS to the MED, our local terminology used by the clinical systems at NYPH. The mapping is not always a straightforward task. The MED may or may not contain the concepts and relations extracted from MEDLINE citations and the UMLS during the KB construction. Manual mapping and editing is necessary.

While medical applications often use KBs to support their reasoning, it is only recently that they have begun to use KBs to support their terminologies [79, 80]. Recent work in the development of knowledge-based representation of terminologies may facilitate the translation of coded data [48]. These techniques enhance the meaning and usability of terminologies [48].

Stead and colleagues define first-generation projects as those that provide integration by using a single system, second-generation as those that integrate data and information across various systems, and third-generation as projects that explicitly relate otherwise separate data and information resources. In third-generation projects, data and knowledge that are outside the system may be linked to the data and processes that are within it [80–86]. Our application of a knowledge-based terminology constitutes an example of the "third-generation," in which the terminology serves as an ontology about relevant relationships among data.

One challenge is the uneven coverage of the terminologies involved in the process. Description of patient conditions and events constitute the basic content of medical records. A study by Chute and colleagues showed that this content is not well represented in the UMLS [45]. The MED content is based on the needs of the NYPH systems. Currently, it does not provide enough coverage for signs and symptoms, for example. We need to determine whether the terminologies and methods that we have chosen are suitable for providing appropriate concept-relationship knowledge and coverage. Finally, there are other concerns related to the specific vocabulary necessary for IR (e.g., fields to search, clinical task, Boolean operators).

Our use of knowledge about the patient to formulate search queries is a significant part of the system. Previous research in our institution [87] has demonstrated that, by using knowledge about the data, we can retrieve, filter, and organize information intelligently, and can reduce the information overload that most clinicians experience. Also important is the potential use of patient data to select the most relevant resources. Patient data could be translated or mapped to a target form that is recognized by the desired information resource. For example, a positive microbiology test shows the presence of an organism and all antibiotic-sensitive test results. Knowing that an antibiotic sensitivity test is linked to a particular pharmaceutical component allows us to use the pharmaceutical component instead of the procedure name to retrieve relevant information in the medical literature.

The implementation of the system raises questions on the generalizability of the methods and model proposed. Will this model be generalizable to guideline databases (e.g., the National Guidelines Clearinghouse), full-text collections of scientific research articles (e.g., PubMED Central [88], BioMed Central [89]), online databases of evidence-based content (e.g., critically appraised topic (CAT)), information in portal sites, or less structured information resources on the Web? The degree to which we can generalize the model will depend on not only the quality of the KB, but also the technical challenges related of integration. The integration of clinical systems and information resources will become easier if resource developers agree to content conventions or standards, such as a defined set of elements, controlled vocabularies, or classification of data elements [13].

V. CONCLUSION

We believe that by helping clinicians to pose well-structured clinical queries, and including in them relevant information from individual patients' medical records, we can enhance IR and thus can improve patient care. We have described the framework for our approach and have delineated significant terminology issues related to the construction of the KB and to the use of patient data to facilitate IR. We are currently evaluating the KB to support information extraction from the electronic medical record and are integrating the extraction and query processes.

ACKNOWLEDGMENTS

We thank Justin Starren and Carol Friedman for their collaboration in the design of the system architecture. We thank Suzanne Bakken

for her support and suggestions. This work was supported by the Center for Advanced Technology of New York State, by Grant IIS-98-17434. Digital Library Initiative 2, National Science Foundation (Kathleen McKeown, principal investigator), and by Grant 20057/95-5, CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil.

REFERENCES

1. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Ann Intern Med 1985; 103(4):596–9.
2. Timpka T, Ekstrom M, Bjurulf P. Information needs and information seeking behavior in primary health care. Scand J Prim Health Care 1989; 7(2):105–9.
3. Shelstad KR, Clevenger FW. Information retrieval patterns and needs among practicing general surgeons: a statewide experience. Bull Med Library Assoc 1996; 84(4):490–7.
4. Jadad AR, Gagliardi A. Rating health information on the Internet: navigating to knowledge or to Babel? JAMA 1998; 279(8):611–4.
5. Gorman PN, Helfand M. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. Med Decis Making 1995; 15(2):113–9.
6. Haynes RB, Hayward RS, Lomas J. Bridges between health care research evidence and clinical practice. J Am Med Inform Assoc 1995; 2(6):342–50.
7. Lindberg DA, Siegel ER, Rapp BA, Wallingford KT, Wilson SR. Use of MEDLINE by physicians for clinical problem solving. JAMA 1993; 269(24):3124–9.
8. Osheroff JA, Bankowitz RA. Physicians' use of computer software in answering clinical questions. Bull Med Library Assoc 1993; 81(1):11–9.
9. Gorman PN, Ash J, Wykoff L. Can primary care physicians' questions be answered using the medical journal literature? Bull Med Library Assoc 1994; 82(2):140–6.
10. Cimino JJ. Linking patient information systems to bibliographic resources. Methods Inf Med 1996; 35(2):122–6.
11. Salton G. Introduction to modern information retrieval. New York, NY: McGraw–Hill, 1983.
12. Hersh WR, Detmer WM, Frisse ME. Information-retrieval systems. 2nd ed. New York: Springer-Verlag, 2001; 539–72.
13. Humphreys BL. Electronic health record meets digital library: a new environment for achieving an old goal. J Am Med Inform Assoc 2000; 7(5):444–52.
14. Forsythe DE, Buchanan BG, Osheroff JA, Miller RA. Expanding the concept of medical information: an observational study of physicians' information needs. Comput Biomed Res 1992; 25:181–200.
15. Lancaster FW, Warner AJ. Information retrieval today. Arlington, VA: Information Resources Press, 1993.
16. Hersh WR. Information retrieval: a health care perspective. In: Orthner HF, editor. Computer and medicine. New York: Springer-Verlag, 1996.
17. Osheroff JA, Forsythe DE, Buchanan BG, Bankowitz RA, Blumenfeld BH, Miller RA. Physicians' information needs: analysis of questions posed during clinical teaching. Ann Intern Med 1991; 114(7):576–81.
18. Ely JW, Osheroff JA, Ebell MH, *et al.* Analysis of questions asked by family doctors regarding patient care. BMJ 1999; 319(7206):358–61.
19. Corcoran-Perry S, Graves J. Supplemental-information-seeking behavior of cardiovascular nurses. Res Nurs Health 1990; 13(2):119–27.
20. Salasin J, Cedar T. Information-seeking behavior in an applied research/service delivery setting. J Am Soc Inf Sci 1985; 36(2): 94–102.
21. Curley SP, Connelly DP, Rich EC. Physicians' use of medical knowledge resources: preliminary theoretical framework and findings. Med Decis Making 1990; 10(4):231–41.
22. Horowitz GL, Jackson JD, Bleich HL. PaperChase. Self-service bibliographic retrieval. JAMA 1983; 250(18):2494–9.
23. Collen MF, Flagle CD. Full-text medical literature retrieval by computer. A pilot test. JAMA 1985; 254(19):2768–74.
24. Markert RJ, Parisi AJ, Barnes HV, *et al.* Medical student, resident, and faculty use of a computerized literature searching system. Bull Med Library Assoc 1989; 77(2):133–8.
25. Haynes RB, McKibbon KA, Walker CJ, Ryan N, Fitzgerald D, Ramsden MF. Online access to MEDLINE in clinical settings. A study of use and usefulness. Ann Intern Med 1990; 112(1):78–84.
26. Abate MA, Shumway JM, Jacknowitz AI. Use of two online services as drug information sources for health professionals. Methods Inf Med 1992; 31(2):153–8.
27. Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. JAMA 1998; 280(15):1347–52.
28. Gorman P. Does the medical literature contain the evidence to answer the questions of primary care physicians? Preliminary findings of a study. In: Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care. 1993.
29. Sackett DL, Straus SE. Finding and applying evidence during clinical rounds: the "evidence cart." JAMA 1998; 280(15):1336–8.
30. Smith R. What clinical information do doctors need? BMJ 1996; 313(7064):1062–8.
31. Medical subject headings—annotated alphabetical list. Bethesda, MD: 1999. [National Library of Medicine]
32. Blois MS, Tuttle MS, Sherertz DD. RECONSIDER: a program for generating differential diagnosis. In: Proceedings of the Fifth Annual Symposium on Computer Applications in Medical Care. IEEE Comput Soc Press, 1981; 3263–8.
33. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain. An evolving diagnostic decision-support system. JAMedA 1987; 258(1):67–74.
34. Miller RA, Masarie FE, Myers JD. Quick medical reference (QMR) for diagnostic assistance. MD Comput 1986; 3(5):34–48.
35. Lowe HJ, Barnett GO. Micro-MeSH: a microcomputer system for searching and exploring the National Library of Medicine's medical subheadings (MeSH) vocabulary. In: Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care. 1987: IEEE Comput Soc Press, 1987; 717–20.

36. Hersh WR, Greenes RA. SAPHIRE—an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. Comput Biomed Res 1990; 23:410–25.

37. Hersh WR, Brown KE, Donohoe LC, Campbell EM, Horacek AE. CliniWeb: managing clinical information on the World Wide Web. J Am Med Inform Assoc 1996; 3(4):273–80.

38. Tuttle MS, Olson NE, Keck KD, *et al.* Metaphrase: an aid to the clinical conceptualization and formalization of patient problems in healthcare enterprises. Methods Inf Med 1998; 37(4–5):373–83.

39. Wingert F. An indexing system for SNOMED. Methods Inf Med 1986; 25(1):22–30.

40. Sherertz DD, Tuttle MS, Blois MS, Erlbaum MS. Intervocabulary mapping within the UMLS: the role of lexical matching. In: Greenes RA, editor. Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. 1988; IEEE Comput Soc Press, 201–6.

41. Elkin PL, Cimino JJ, Lowe HJ. Mapping to MeSH (the art of trapping MeSH equivalence from within narrative text). In: Greenes RA, editor. Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. IEEE Comput Soc Press, 1988; 185–90.

42. Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. MD Computing 1990; 7(2):104–9.

43. Masarie FE Jr, Miller RA, Bouhaddou O, Giuse NB, Warner HR. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res 1991; 24:379–400.

44. Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med 1993; 32(4):281–91.

45. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. J Am Med Inform Assoc 1996; 3(3):224–33.

46. Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures. J Am Med Inform Assoc 1997; 4(3):238–51.

47. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998; 37(4–5):394–403.

48. Bakken S, Cashen MS, Mendonça EA, O'Brien A, Zieniewicz J. Representing nursing activities within a concept-oriented terminological system: evaluation of a type definition. J Am Med Inform Assoc 2000; 7(1):81–90.

49. Campbell KE, Cohn SP, Shortliffe EH, Rennels G. Scalable methodologies for distributed development of logic-based convergent medical terminology. Methods Inf Med 1998; 37(4–5):426–39.

50. Rector AL, Bechhofer S, Goble CA, Horrocks I, Nowlan WA, Solomon WD. The GRAIL concept modelling language for medical terminology. Artif Intell Med 1997; 9(2):139–71.

51. McKibbon A, Eady A, Marks S. PDQ evidence-based principals and practice. Hamilton, Ontario, BC: Dekker, 1999. [PDQ Series]

52. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. JAMA 1994; 1(6):447–58.

53. Hersh W. "A world of knowledge at your fingertips": the promise, reality, and future directions of on-line information retrieval. Acad Med 1999; 74(3):240–3.

54. Loonsk JW, Lively R, TinHan E, Litt H. Implementing the medical desktop: tools for the integration of independent information resources. In: Clayton PD, editor. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care. 1991; 574–7.

55. Powsner SM, Miller PL. From patient reports to bibliographic retrieval: a Meta-1 front-end. Proc Annu Symp Comput Appl Med Care 1991; 526–30.

56. Powsner SM, Riely CA, Barwick KW, Morrow JS, Miller PL. Automated bibliographic retrieval based on current topics in hepatology: hepatopix. Comput Biomed Res 1989; 22:552–64.

57. Powsner SM, Miller PL. Automated online transition from the medical record to the psychiatric literature. Methods Inf Med 1992; 31(3):169–74.

58. Cooper GF, Miller RA. An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text. J Am Med Inform Assoc 1998; 5(1):62–75.

59. Cimino C, Barnett GO, Laboratory of Computer Science–Massachusetts General Hospital. Standardizing access to computer-based medical resources. In: Miller RA, editor. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care. Washington, DC: 1990; 33–7.

60. Cimino JJ, Johnson SB, Aguirre A, Roderer N, Clayton PD. The MEDLINE button. Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care. 1992; 81–5.

61. Cimino JJ, Elhanan G, Zeng Q. Supporting infobuttons with terminological knowledge. In: Masys DR, editor. Proceedings/AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus, 1997; 528–32.

62. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996; 312(7023):71–2.

63. Friedland DJ, Go AS, Davoren JB, *et al.* Evidence-based medicine. A framework for clinical practice. Stamford, CT: Appleton & Lange, 1998.

64. Hunt DL, Haynes RB, Browman GP. Searching the medical literature for the best evidence to solve clinical questions. Ann Oncol 1998; 9(4):377–83.

65. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. 2nd ed. London: Churchill Livingstone, 2000.

66. Cimino JJ, Aguirre A, Johnson SB, Peng P. Generic queries for meeting clinical information needs. Bull Med Library Assoc 1993; 81(2):195–206.

67. Commission on Professional and Hospital Activities. International classification of diseases. 9th rev with clinical modifications (ICD9-CM). Ann Arbor: 1978.

68. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc 1994; 1(1):35–50.

69. Cimino JJ, Sengupta S, Clayton PD, Patel VL, Kushniruk A, Huang X. Architecture for a Web-based clinical information system that keeps the design open and the access closed. In: Chute CG, editor.

Proceedings/AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus, 1998; 121–5.

70. Hripcsak G, Cimino JJ, Sengupta S. WebCIS: large scale deployment of a Web-based clinical information system. Proceedings/AMIA Annual Fall Symposium. 1999; 804–8.
71. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994; 1(2):161–74.
72. Mendonça EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. Proc AMIA Symp 2000; (20 Suppl):575–9.
73. Sowa JF. Knowledge representation—logical, philosophical, and computational foundations. Pacific Grove, CA: Brooks/Cole, 2000.
74. Seol YH, Johnson SB, Cimino JJ. Conceptual guidance in information retrieval. Submitted. 2001.
75. Ely JW, Osheroff JA, Gorman PN, *et al.* A taxonomy of generic clinical questions: classification study. BMJ 2000; 321(7258): 429–32.
76. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. New York: Churchill Livingstone, 1997.
77. Fletcher RH, Fletcher SW. What is the future of internal medicine? Ann Intern Medic 1993; 119(11):1144–5.
78. Bakken S. An informatics infrastructure is essential for evidence-based practice. J Am Med Inform Assoc 2001; 8(3):199–201.
79. Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the medical entities dictionary. J Am Med Inform Assoc 2000; 7(3):288–97.
80. Spackman KA, Campbell KE, Côté RA, authors. SNOMED RT: a reference terminology for health care. In: Masys DR, editor. Proceedings/AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus, Inc., 1997; 640–4.
81. Stead WW, Miller RA, Musen MA, Hersh WR. Integration and beyond: linking information from disparate sources and into workflow. J Am Med Inform Assoc 2000; 7(2):135–45.
82. Campbell KE, Musen MA. Representation of clinical data using SNOMED III and conceptual graphs. Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care. 1993; 354–8.
83. Campbell KE, Cohn SP, Chute CG, Rennels G, Shortliffe EH. Galapagos: computer-based support for evolution of a convergent medical terminology. Proceedings/AMIA Annual Fall Symposium. 1996; 269–73.
84. Rector AL, Nowlan WA, Glowinski A. Goals for concept representation in the GALEN project. Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care. 1994; 414–8.
85. Brown PJ, O'Neil M, Price C. Semantic definition of disorders in version 3 of the read codes. Methods Inf Med 1998; 37(4–5): 415–9.
86. Hardiker NR, Rector AL. Modeling nursing terminology using the GRAIL representation language. J Am Med Inform Assoc 1998; 5(1):120–8.
87. Zeng Q, Cimino JJ. Evaluation of a system to identify relevant patient information and its impact on clinical information retrieval. Proc AMIA Symp 1999; 642–6.
88. PubMED Central [Web Page]. Available at http://www.pubmedcentral.nih.gov. [accessed 17 May 2001]
89. BioMED Central [Web Page]. Available at http://www.biomedcentral.com/. [accessed 17 May 2001]
90. Zeng Q. A knowledge-based concept-oriented view generation system for clinical data [dissertation]. New York (NY): Columbia University, 1999.
91. Elhadad N, McKeown K. Towards generating patient specific summaries of medical articles. Proceedings of NAACL Workshop on Automatic Summarization, 2001, in press.
92. Kan MY, McKeown K, Klavans JL. Applying natural language generation to indicative summarization. Proceedings of the 8th European Workshop on Natural Language Generation, 2001, in press.